

Om aber og 02-taller: Præciseringer og refleksioner



Peter Sunde,
Institut for Bioscience,
Aarhus Universitet



Pernille Bødtker Sunde,
DPU, Aarhus Universitet,
og VIA Læreruddannelsen
i Aarhus

Kommentar til kommentarerne i MONA, 2017-1, til Sunde & Sunde: "Den smarte abe: betydning af og korrektion for gætning ved karaktergivning i multiple choice-tests", MONA, 2016-4.

Principper og praksis for karaktergivning efter 7-trinsskalaen og benyttelse af multiple choice-tests (MCT's) som eksamensredskab synes at være udtømmelige debattemner i faglige såvel som politiske fora. I så henseende er det ikke overraskende at vores artikel om betydning af gætning for point- og karaktergivning ved MCT (Sunde & Sunde, 2016) affødte tre forskellige kommentarer (Lotte Dyhrberg O'Neill, 2017 [LDO'N]; Svend Kreiner, 2017 [SK]; Peter Weng, 2017 [PW]) vedrørende de specifikke resultater fra vores analyse og MCT som evalueringsredskab i almindelighed. I det følgende har vi forsøgt at samle op på de dele af kommentarerne som vedrørte vores analyse og konklusioner.

For det første ønsker vi at præcisere mulige misforståelser vedrørende brug af nulfordelinger til at kvantificere hvor mange spørgsmål man reelt skal kunne kende svarene på for at opnå en given rigtig svarprocent i eksamener baseret på MCT's. Til dette hører også spørgsmålet om hvorvidt gætning er et udpræget fænomen i virkeligheden, og om manglende korrektion for gætning udgør et reelt problem for karakterfastsættelsen.

For det andet mener vi at især SK's meget oplysende indlæg om Ministeriet for Børn, Unge og Ligestillings (UVM's) principper og praksis for karaktergivning efter 7-trinsskalaen i folkeskolens afgangsprøve for biologi fortjener opmærksomhed idet det giver anledning til at rette fokus på den inkonsistens som tilsyneladende råder omkring principper for karaktergivning efter 7-trinsskalaen. Som man kan læse af SK's indlæg, er UVM's karaktergivning efter 7-trinsskalaen i højere grad norm- end kriteriebaseret, modsat hvad man kan forledes til at tro ud fra læsning af UVM's egen bekendtgørelse (UVM 2007) og senest opdaterede instruks (UVM 2017).

Anvendelse af nul-modeller til beregning af pointinflationspotentialet af gætning

LDO'N og SK opponerer i deres kommentarer mod brugen af nul-modeller til at estimere hvor mange rigtige svar en eksaminand kunne forventes at opnå gennem blind gætning i tillæg til de spørgsmål/items som han/hun rent faktisk kan svaret på. De kritiserer begge den kvantitative nul-model ("abemodellen") for at være utilstrækkelig til at beskrive elevers virkelige responsadfærd hvor der typisk vil være alle mulige overgange fra det sikre svar til det mere eller mindre kvalificerede eller ukvalificerede gæt (citerer: "*Der er forskel på elever og aber*" og "*den abe-adfærd modellen forudsætter, har meget lidt med virkeligheden at gøre*". SK, s. 80).

Vi er helt enige i at den anvendte nul-model med kun to responsudfald (sikkert svar vs. blindt gæt) er en grov forsimpning af den virkelige elevadfærd idet (som vi selv skrev) "*der i virkelighedens verden er en glidende overgang fra det stensikre svar givet uden tvivl over det kvalificerede gæt til det helt ukvalificerede gæt*" (Sunde & Sunde, 2017, s. 33). At en kvantitativ model er en forsimpning af virkeligheden og dermed per definition er forkert (det er det der definerer en model), betyder imidlertid ikke at dens resultater ikke er anvendelige (Box & Draper, 1987). I så henseende skal resultaterne fra "abe"-modellen betragtes som referenceværdier (pejlemærker) for hvor stor en effekt ren gætning (dvs. den mest ekstreme form for mangel på faglig kunnen) har for pointopnåelse i et testresultat. I så henseende vil vi fastholde at "abe"-modellens resultater er konservative i forhold til virkelighedens verden hvor en stor del af de målrettede svar ikke er afgivet på et stensikkert, men derimod på mere eller mindre kvalificeret vidensgrundlag (dvs. at eleven ikke er 100 % sikker, og at svarene dermed indeholder et element af gæt) som også redegjort for af SK. Vi vil dermed fastholde konklusionen fra vores analyse, at en elev for at have 50 % chance for at bestå folkeskolens afgangsprøve i biologi i maj 2005 som minimum behøvede at have en paratviden svarende til at kunne kende svaret på 17 % af alle spørgsmål med 100 % sikkerhed for at opnå 36 % rigtige svar og dermed et 02-tal.

Hvorfor skulle gætning ikke forekomme i virkeligheden?

Både LDO'N og SK argumenterer med at danske børn kun i ringe grad gætter svar på spørgsmål de ikke kan svaret på. Som evidensgrundlag refererer LDO'N i den forbindelse til en analyse af Nørgaard et al. (2006). Vi vil her gøre opmærksom på at Nørgaard et al.s (2006) konklusion efterfølgende blev modsagt af Allerup (2006).

I mangel af egentlig evidens for eller imod forekomst af gætning i prøver hvor eleverne er usikre på svaret, vil vi vove den påstand at det er direkte ulogisk at gætning ikke er vidt udbredt som en naturlig og hensigtsmæssig problemløsningsstrategi i en

test hvor forkerte svar ikke straffes (hvilket vi endnu en gang pointerer heller ikke bør forekomme da dette fører til at børn underpræsterer ved at afholde sig fra at svare på spørgsmål de ikke er helt sikre på er rigtige: Betts et al., 2009). Som redegjort for i det ovenstående ligger der i selve afgivelsen af et ikke helt sikkert svar et element af gætning. At tage chancen på et svar man ikke er sikker på, er desuden intuitivt og reelt den bedste løsningsstrategi. Dette er også den entydige erfaring den ene af os (PBS) har gjort sig gennem 11 år som lærer i folkeskolen. I forbindelse med researchen til vores artikel blev dette indtryk bekræftet af en række lærere som fortsat arbejder i folkeskolen.

Lad os derfor hellere erkende og anerkende at i forbindelse med MCT's er gætning en naturlig og legitim del af spillet. Konsekvensen af ikke at instruere ALLE børn i at man altid skal afgive et svar, også når man er usikker, er at man systematisk handicapper de tilbageholdende og autoritetstro børn i forhold til de risikovillige "regelbrydere". PW's refleksion over hvorvidt bedre testresultater for drenge end for piger i naturfaglige tests delvist kan forklares med en højere risikovillighed hos drenge end hos piger, kunne således være interessant at få undersøgt nærmere.

Der er andre og mere betydende usikkerhedskilder end gætning i forbindelse med MCT's.

LDO'N, SK og PW fremhæver alle at gætning kun er én ud af mange usikkerhedsmomenter i forbindelse med tilrettelæggelse af MCT's. Af andre fejlkilder nævnes relevansen af de stillede spørgsmål og betydningen af tilfældig usikkerhed omkring hvilke af de udvalgte spørgsmål som en eksaminand kan svare på. Vi anerkender til fulde disse øvrige usikkerhedsmomenter som er forbundet med MCT's (og eksamenssæt i almindelighed), som vi for klarhedens skyld ikke inddrog i vores artikel. At der også måtte være andre problemer med MCT's som eksamensform, ændrer dog ikke grundlæggende ved vores pointer vedrørende gætning og risikovillige svar. For ordens skyld vil vi endnu en gang pointere at vores ærinde ikke var at kritisere MCT som eksamensform, men blot gøre opmærksom på hvilken betydning gætning kan have for pointtildeling og kriteriebaseret karaktergivning.

Bliver 7-trinsskalaen i virkeligheden brugt normbaseret?

I sin kommentar angiver SK (som har hjulpet UVM med at for-censurere prøverne i 2015 og 2016) at karaktergivningen i UVM's centralt stillede prøver slet ikke er kriteriebaseret som angivet i UVM's karakterbekendtgørelse (2007), hvilket var en del af præmissen for vores artikel. SK (s. 82) kan således oplyse at "*Da karakterskalaen blev indført i 2007, var der, på trods af ministeriets snak om kriteriebaserede karakterer, tale*

om et normbaseret karaktersystem". I forbindelse med censureringen af de seneste års opgaver kan SK desuden oplyse at "det der har været centralt, har været at sikre at karaktererne i forbindelse med den seneste prøve svarer til karaktererne fra tidligere år. De var normbaserede i 2007, men man har siden forsøgt at fastlægge karaktererne så de bliver sammenlignelige hen over tiden". SK konkluderer selv at: "Af den grund er karaktererne hverken fugl eller fisk. Hverken normbaserede eller kriteriebaserede". SK skriver endvidere i sin forklaring (s. 81) at "det er let at se at ministeriet ubetænksomt taler med to tunger, og at der er tale om karakterer der er tættere på at være normbaserede end kriteriebaserede".

Dette er befriende ærlig tale som vi gerne vil give SK en stor cadeau for at bringe frem i lyset. Som SK indirekte skriver, må man heraf konkludere at der tilsyneladende ikke er overensstemmelse imellem de principper efter hvilke 7-trinsskalaen er defineret, og den måde den konkret anvendes på. Dette kan enhver forsikre sig om ved at læse UVM's seneste instruks om brug af 7-trinsskalaen som er dateret 13. marts 2017 (<http://www.uvm.dk/uddannelsessystemet/7-trins-skalaen/anvendelse-af-7-trins-skalaen>). Her fremgår det nemlig at (vores understregning): "Karakterskalaen skal anvendes absolut. Nationalt og over en årrække kan skalaen på denne måde siges at være relativ. Men i den konkrete brug er det afgørende, at karakterskalaen anvendes absolut med udgangspunkt i beskrivelserne af de enkelte karakterer og i forhold til fagets og uddannelsens mål. Den enkelte præstation skal derfor bedømmes ud fra karakterbeskrivelserne i karakterbekendtgørelsen set i forhold til målbeskrivelsen for uddannelsesforløbet." Dette læser vi (andre kan have en anden tolkning) således at det officielle princip fortsat synes at være at karakterer skal fastlægges absolut efter konkrete læringsmål. Læringsmålene skal så fastlægges så fordelingen af karaktererne stemmer overens med fraktionsfordelingen ifølge ECTS-skalaen. Praksis (jf. SK) synes dog at være at karaktererne inden for hvert enkelt år gives efter en ren normfordeling. Så ved vi det.

Refleksion: Hvad er de præcise faglige krav til at opnå et 02-tal?

Uanset om karakterfastsættelsen i folkeskolens afgangsprøve i biologi specifikt (og eksamener i det danske uddannelsessystem generelt) er kriteriebaseret eller normbaseret i tråd med ECTS-skalaen, står et ret principielt spørgsmål fra vores artikel fortsat ubesvaret tilbage: Hvad er det præcise faglige krav til at opnå den laveste beståelses-karakter, 02 (den tilstrækkelige præstation)? Normfordelingen for karaktergivning efter ECTS-skalaen (Europakommissionen, 2015) omfatter nemlig kun de *beståede* elever. Heri følger at selv efter ECTS-skalaen skal de faglige målopfyldelseskriterier for opnåelse af beståelseskarakteren fremgå.

I vores artikel beregnede vi ud fra UVM's omsætningstabel for karaktergivning det omtrentlige faglige beståelseskriterium for folkeskolens afgangsprøve i biologi, maj 2015, til at svare til at eksaminanden skulle have stensikker viden om ca. 17 % af de adspurgte spørgsmål (hvilket med brug af gætning ville føre til ca. 36 % rigtige svar i testen). Hvilke forudgående faglige overvejelser der har ligget til grund for fastsættelsen af lige præcis 36 % rigtige svar som målopfyldelseskriterium for karakteren O2, skal vi lade være usagt. Men nu kan alle forholde sig til at det svarer til ca. 17 % sikker viden i forhold til de adspurgte spørgsmål inden for det emnevalg og med den sværhedsgrad som blev valgt det pågældende år.

Men måske man skulle starte et helt andet sted, nemlig ved at gøre det obligatorisk for alle eksamener at definere de faglige målopfyldelseskriterier for O2-taller i absolutte termer på samme måde som det nu gøres for 12-tallet? Når det kommer til stykket, er O2-tallet nemlig en langt vigtigere karakter end 12-tallet fordi O2-tallet definerer den tilstrækkelige præstation og dermed adgangsbilletten til embede og ansvar, mens 12-tallet er en luksuskarakter som blot fortæller at præstationen opfyldte alle fagets læringsmål.

Referencer:

- Allerup, P. (2006). At trykke på "statistikknappen" er ikke tilstrækkeligt. *MONA*, 2006(4), s. 67-69.
- Betts, L.R., Elder, T.J., Hartley, J. & Trueman, M. (2009). Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes? *Assessment & Evaluation in Higher Education*, 34(1), s. 1-15.
- Box, G.E.P & Draper, N.R. (1987). *Empirical model-building and response surfaces* (Vol. 424): Wiley New York.
- Europakommissionen (2015). ECTS User's guide: Grade distribution http://ec.europa.eu/education/ects/users-guide/grade-distribution_en.htm. Lokaliseret 4. april 2017.
- Kreiner, S. (2017). Aber eller elever: en kommentar. *MONA*, 2017(1), s. 80-84.
- Nørgaard, K. (2006). De digitale afgangsprøver har høj kvalitet. *MONA*, 2006(3), s. 86-92.
- Ministeriet for Børn, Unge og Ligestilling (2007). Bekendtgørelse om karakterskala og anden bedømmelse. *BEK nr 262 af 20/03/2007*. <https://www.retsinformation.dk/Forms/R0710.aspx?id=25308>. Lokaliseret 4. april 2017.
- Ministeriet for Børn, Unge og Ligestilling (2017). Anvendelse af 7-trinsskalaen. <http://www.uvm.dk/uddannelsessystemet/7-trins-skalaen/anvendelse-af-7-trins-skalaen>. Lokaliseret 4. april 2017.
- O'Neill, L.D. (2017). Aben der nægtede at dø. *MONA*, 2017(1), s. 74-79.
- Sunde, P. & Sunde, P.B. (2016). Den smarte abe. *MONA*, 2016(4), s. 24-35.
- Weng, P. (2017). Skal vi bruge krudt på at skyde en "smart abe" ned? *MONA*, 2017(1), s. 85-88.